LARS Information Note 062276

TECHNIQUES AND APPLICATIONS FOR COMPUTER-AIDED ANALYSIS OF MULTISPECTRAL SCANNER DATA

BY

ROGER M. HOFFER

The Laboratory for Applications of Remote Sensing

Purdue University, West Lafayette, Indiana

1976

LARS Information Note 062276

TECHNIQUES AND APPLICATIONS FOR COMPUTER-AIDED ANALYSIS OF MULTISPECTRAL SCANNER DATA

by

Roger M. Hoffer

An Invited Paper Presented at the
XVI World Congress of the
International Union of Forest Research Organizations
Oslo, Norway
June 1976

TECHNIQUES AND APPLICATIONS FOR COMPUTER-AIDED ANALYSIS OF MULTISPECTRAL SCANNER DATA

Roger M. Hoffer, Professor of Forestry

Purdue University, West Lafayette, Indiana, U.S.A.

ABSTRACT

Several procedures for digitally processing and analyzing data from satellite scanner systems that have been found to be particularly useful are described. The techniques were applied to a mountainous test site of approximately one million hectares in area. In spite of the vegetative and topographic complexity of this test site, coniferous and deciduous forest cover, as well as other major cover types, could be mapped with an accuracy of approximately 85% using both LANDSAT and SKYLAB data. Individual forest cover types were mapped with approximately 70% accuracy. Accurate acreage estimates of forest cover were obtained through use of these techniques over large geographic areas.

INTRODUCTION

In forestry, as in many other discipline areas, there exists a distinct need for timely, reliable information concerning the resource base with which one is working. Because of the extensive nature of the world's forest resources, the synoptic type of data that can be obtained from spacecraft altitudes has been of considerable interest during the past few years. LANDSAT-1 (previously ERTS-1) clearly showed that high quality data can be obtained at frequent intervals over most of the earth's surface. As the value and the potential of LANDSAT data has become more apparent, many questions are being raised concerning techniques that can most effectively handle and analyze such masses of data. One approach, which was first attempted in 1966 at the Laboratory for Applications of Remote Sensing (LARS), Purdue University, involves the use of pattern recognition techniques applied to measurements obtained from multispectral scanner (MSS) systems. This approach was initially developed for agricultural situations and used data obtained from aircraft altitudes. In the last three years, modification and refinement of the basic procedures and extensive testing with LANDSAT and SKYLAB scanner data has

---

The research reported in this paper was supported by NASA Contracts NAS9-13380, NAS5-21880 and NASA Grant No. NGL 15-005-112.

proven that these computer-aided analysis techniques can be successfully utilized for mapping wildland natural resources from spacecraft altitudes.

It is the purpose of this paper to briefly describe some of the techniques utilized in computer-aided analysis of satellite MSS data and to report the results of recent work with LANDSAT and SKYLAB data for mapping forest cover.

COMPUTER-AIDED ANALYSIS TECHNIQUES

Current procedures for digitally processing and analyzing data from LANDSAT, SKYLAB, or other multispectral scanner systems involve four primary areas of activity. These include data reformatting and preprocessing, analysis and classification of the data, information display and tabulation, and finally, evaluation of the results. The data reformatting and preprocessing involves such activities as reformatting LANDSAT scanner data to allow a full frame to be contained on a single data tape, digital filtering to improve data quality (signal-to-noise ratio), geometrically correcting and scaling the data to a common map base, and digital registration of multiple sets of scanner and other data. Such procedures do not involve any actual analysis of the data, but simply allow subsequent data analysis activities to be carried out in a much more effective manner.

The analysis and classification of MSS data involves a series of steps designed to enable the computer to identify various cover types or earth surface features of interest. The key element in such computer-aided analysis techniques involves a man/machine interaction, whereby the man "trains" the computer to recognize particular combinations of numbers that represent reflectance measurements in each of several wavelength bands, for the particular cover types of interest. This training process is carried out utilizing scanner data collected over a limited geographic area. Then, after a good set of training statistics have been developed, the computer is programmed to classify the reflectance values for each resolution element in the entire data set. In this way, the computer can map and tabulate cover types over a large geographic area at a much faster rate than would be possible for a man using standard image interpretation techniques.

One of the first considerations in computer-aided analysis of multispectral data involves the definition of the categories or classes of material that the computer should be trained to recognize. Basically, there are two conditions which must be met by each class involved in an analysis of multispectral scanner data using these computer-aided analysis techniques:

- The class must be spectrally separable from all other classes.
- The class must be of interest to the user or have informational value.

In working with multispectral scanner data, one often finds that the classes of interest to the user cannot be spectrally separated at certain times of the year. Quite often, different species of green vegetation have very similar spectral characteristics, even though their morphological charac-

teristics may be quite different. The need for a class to be both separable and to have informational value leads to two quite different basic approaches in training the computer system.

The first approach is referred to as the "supervised technique", and involves use of a system of X-Y coordinates to designate to the computer system the locations of known earth surface features that have informational value. For example, at a certain X-Y location in the data is a stand of ponderosa pine; another location is a stand of aspen; other areas contain Douglas fir, grassland, water, etc. This supervised technique has been used quite effectively for agricultural mapping, but experience has shown that for wildland areas, where the cover types of interest are not as spectrally homogeneous, this supervised technique often does not enable adequate accuracy or reliability to be achieved. The primary reason for this is the difficulty in defining locations in the data that are representative of all variations in spectral response for every cover type of interest.

A second approach to training the computer system involves the "clustering" technique (sometimes referred to as the "non-supervised" technique). In this approach the analyst simply designates the number of spectrally distinct classes into which the data to be classified should be divided. The computer is programmed to classify the data into the designated number of spectral classes and then prints out a map indicating which resolution elements in the data belong to which spectral classes. The analyst then relates this classification output map to known surface observation data, and determines which materials are represented by each of the different spectral classes indicated on the map (e.g. Spectral Class 1 is aspen, Class 2 is ponderosa pine, etc.). The problem with this technique is that the analyst doesn't know for sure how many spectral classes are actually present. Also one often finds that the classes of most interest have subtle spectral differences while many of the other classes present in the data may be easily separated spectrally but are of little informational value. In spite of these difficulties, much of the early work with this clustering technique indicated that it was much better than the "supervised" technique when working in wildland or natural areas. With the advent of LANDSAT-1, computer-aided mapping of relatively large geographic areas became more feasible. It was found, however, that when large wildland areas were to be analyzed, the amount of data and the number of spectral classes involved became too large to allow the clustering technique to be effectively utilized.

A so-called "modified clustering" technique was therefore developed, and has proven to be extremely effective in working with satellite multispectral scanner data, both from the LANDSAT and SKYLAB scanner systems (Fleming, 1974; Hoffer, 1975). This technique involves a combination of both the clustering and the supervised approaches. In this method, several small blocks of data are defined, each of which contains several cover types (Figure 1). Each area or data block is first clustered separately, and the spectral classes for all cluster areas are subsequently combined. In essence, the modified cluster approach entails discovering the natural groupings present in the scanner data, and then correlating the resultant spectral classes with the desired informational classes (cover types, vegetative conditions, etc.). Normally, less than one percent of the data involved in the final analysis is used for the training phase.

After the training statistics are defined, the maximum likelihood algorithm is utilized in a supervised mode of operation to classify the entire data set. This is a relatively simple pattern recognition algorithm, and has been used successfully at LARS and elsewhere for analysis of remote sensor data involving many different types of applications. In the classification sequence, each resolution element sensed by the scanner system is assigned to one of the spectral classes defined during the development of the training statistics. These classification results are then stored on magnetic tape, and the analyst can subsequently display these results in a variety of map or tabular output formats.

Computer classification results can be displayed in a maplike format using either of two basic techniques. The first involves a logogrammatic printout obtained from a standard computer line printer, in which the analyst selects various symbols to represent each of the different cover types of interest such as D for deciduous forest, C for coniferous forest, W for water, etc., such as

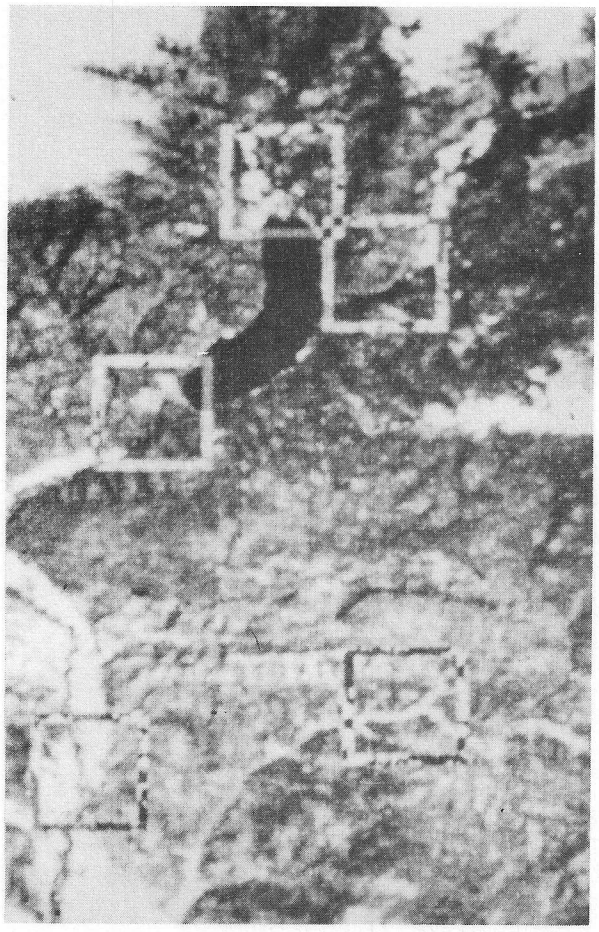

Figure 1 -- Digital Display Example of LANDSAT Data Illustrating the "Modified Cluster" Technique.

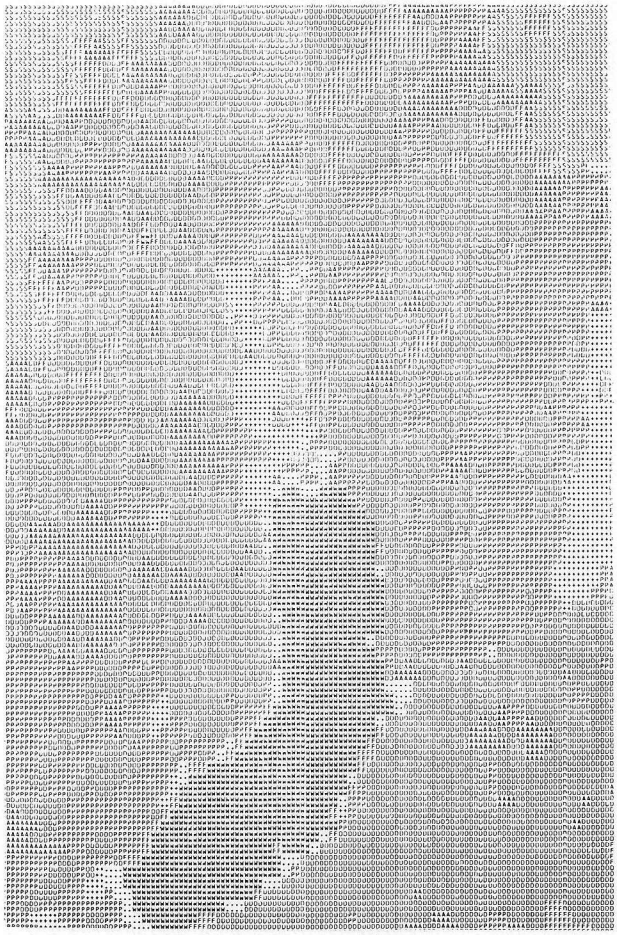

Figure 2 -- Line-Printer Display of Classification Results Using SKYLAB Scanner Data. Key: P=Ponderosa pine, D=Douglas & white fir, F=Engelmann spruce & Subalpine fir, A=Aspen, O=Oak, S=Snow, W=Water, +=Grassland, and .=Exposed Rock & Soil.

shown in Figure 2. A second type of map output can be obtained through the use of a digital display device in which each of the different classes of interest are displayed as a different color or as a different tone of grey, depending on the capabilities of the display. Such display devices provide a much more photographic-like output of the classification results (as compared to the line-printer output) and also have the advantage of displaying much larger geographic areas on a relatively small image (as can be observed by comparing Figures 1 and 2).

Tabular outputs of classification results can be obtained easily, and are particularly useful for acreage determinations over any particular area of interest. In this case, the analyst simply designates to the computer the X-Y coordinates representing the boundary of a test area (such as a quadrangle, a county, or a watershed). The computer then summarizes the number of data points classified into each of the various cover type categories. Since each data point or resolution element of satellite data represents a certain area on the ground (approximately 0.46 hectares per resolution element for LANDSAT), a conversion factor is applied to determine the number of hectares of each of the various cover types of interest. The percentage of the entire area covered by each of the cover types of interest can also be rapidly calculated.

One can also tabulate the classification results from small "test areas" of known cover types. The X-Y coordinates of a statistical sample of test areas of known cover types are designated and the cover types into which these test areas were classified are then tabulated by the computer. These results are than compared to the cover type known to be actually present on the ground, thereby enabling the analyst to quantitatively determine the classification performance by the computer. An example of this type of output is shown later in Table 1.

Classification of large geographic areas can be accomplished very rapidly using computer analysis techniques. However, one must be able to verify the accuracy of such computer classification results. Are the resultant classification maps and tables reasonably accurate, and do they have a reasonable degree of reliability? Several different techniques have been developed and utilized to evaluate such computer classification results. Our experience at LARS has been that a combination of three different techniques provides the best overall indication of the classification accuracy. A qualitative evaluation of the classification results can be obtained by visually comparing the classification to an existing cover type map or to aerial photos of the region. Although the method is subjective, it does provide a quick, rough estimate of the accuracy of the classification. However, quantitative evaluation techniques allow more definitive evaluations of the computer classification results to be obtained.

One quantitative evaluation technique involves a sample of individual areas of known cover types which are designated as "test areas", as discussed previously. To avoid any possible bias on the part of the analyst, the test areas should be located prior to the classification and should be located by means of a statistical sampling design (as illustrated in Figure 3). The cover type classification obtained by the computer for the various test areas is tabulated and compared to the actual cover type present in the test areas. This tabulation can involve the individual test areas or can be summarized for the entire set of test areas.

A second quantitative method of evaluating the computer classification results is to compare acreage estimates obtained from the computer classification of satellite data to those obtained by some conventional method, such as manual interpretation of aerial photos. If an adequate number of relatively large areas are summarized, a statistical correlation can be obtained, thereby enabling a quantitative comparison between the computer-derived acreage estimates and the acreage estimates derived from conventional sources. Figure 4 is an illustration of such a comparison involving a classification of SKYLAB data.

There are many variations and refinements that can be incorporated into these analysis and evaluation techniques. However, the general approaches described above have been found to be most effective for computer-aided mapping of general land use and forest cover types, utilizing either LANDSAT or SKYLAB multispectral scanner data.

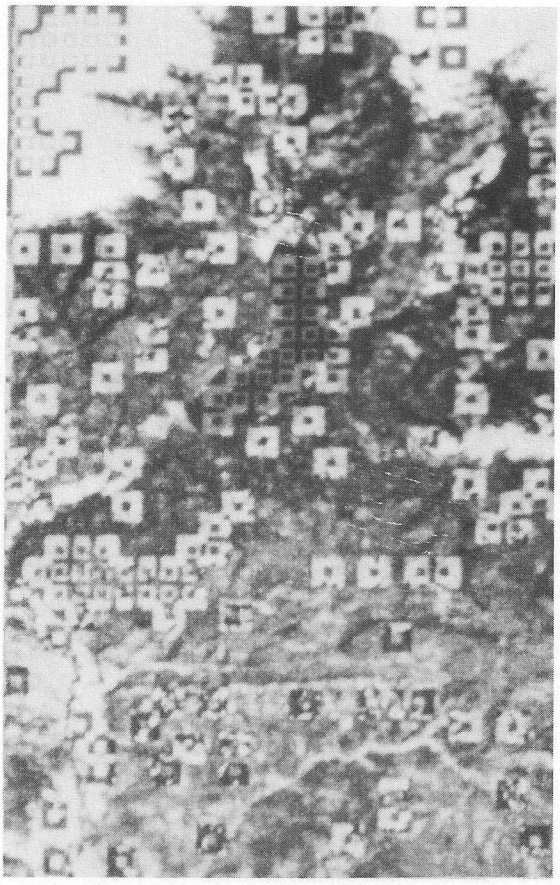

Figure 3 -- Illustration of Statistical Sample of Test Areas.

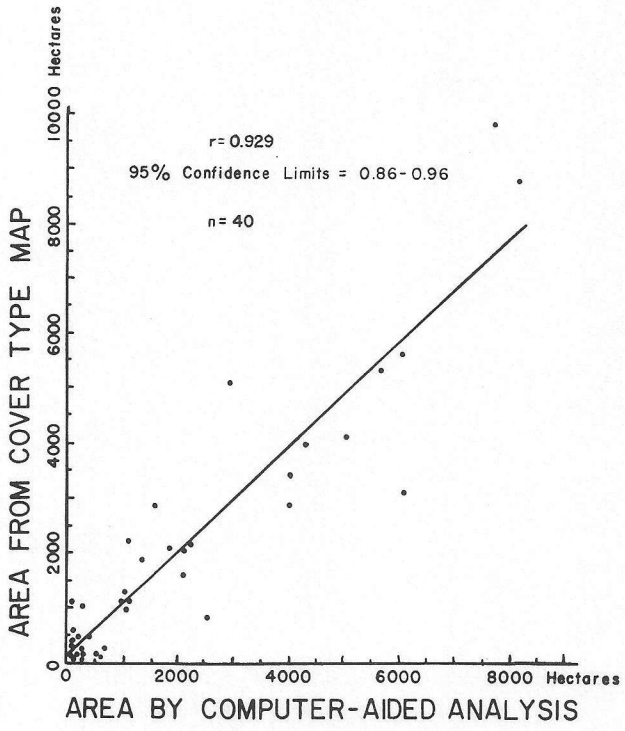

Figure 4 -- Acreage Estimates of Cover Types Obtained by Computer-Aided Analysis of SKYLAB MSS Data, Compared to Aerial Photo Interpretation Estimates.

CLASSIFICATION RESULTS

The computer-aided analysis and evaluation techniques described above were utilized on LANDSAT and SKYLAB data obtained over a mountainous test site in the San Juan Mountains of Southwest Colorado. The area contains a complex mixture of forest types, rangeland, alpine tundra, agricultural areas, water bodies, geological features and various man-made features. The topography of the test site area is rugged, ranging in elevation from less then 2000 meters to over 4200 meters. Within this range of elevation, there is a distinct distribution of cover types according to altitude. Much of the area is dominated by Ponderosa pine (Pinus ponderosa) forest, but Douglas fir (Pseudotsuga menziesii var. glauca), Engelmann spruce (Picea engelmanii), and subalpine fir (Abies lasiocarpa) are found at the higher elevations and on steep north slopes. There are also many stands of quaking aspen (Populus tremuloides), primarily on sites that have been disturbed by fire or avalanches. At lower elevations, the drier steep southern slopes are dominated by Gambel oak (Quercus gambelii), and the valley bottoms are occupied by agricultural land (mostly hayfields). Timberline in the region is approximately 3600 meters, and extensive areas of tundra are found above this elevation.

Several studies involving land use and forest cover mapping were conducted in this test site. One of the major studies utilized LANDSAT data to classify the entire San Juan Mountain Test site, an area of 993,800 hectares (2,456,000 acres). The modified cluster technique was utilized with sixteen training areas (Figure 1), each of which contained four to six cover types. Each training area was clustered into 12 to 18 spectral classes and spectrally similar classes were combined, resulting in 14 distinct, separable spectral classes. Each of these spectral classes were identified using existing aerial photography and type maps of the area. It was determined that the spectral classes present could be grouped into five "Major Cover Types" of Level II "Land Use" Categories[1/], including Coniferous Forest, Deciduous Forest, Grassland, Water, and Barren (exposed rock outcrops, soil, and sparsely vegetated tundra). The grassland category included both cultivated pasture and rangeland areas because they could not be spectrally separated on a reliable basis.

Test areas which included a total of 16,170 resolution elements were then obtained from quadrangles in which no training areas were located. Aerial photos and subsequent field checks were used to identify the characteristics of these test areas. After the entire area was classified, qualitative evaluation of the resultant maps indicated that the classification was reasonably accurate. To obtain a quantitative evaluation, the results in the test areas were tabulated, as shown in Table 1.

This type of tabular display allows an effective method to evaluate both the inclusive and exclusive errors present in the classification, and to determine performance for individual cover types, as well as for the overall classification. Table 1 indicates that a relatively accurate classification had been

---

[1/] Based upon the Level II Land Use Categories, as defined by U.S. Geological Survey Circular 671 (Anderson, 1972).

obtained with these computer-aided analysis techniques. This result was believed to be particularly significant in view of the topographic and vegetative complexity of the area, and the size of the test site involved.

To evaluate the classification using acreage comparisons, the number of resolution elements classified into each cover type within each of the 63 quadrangles (U.S.G.S. 7-1/2 minute quadrangles) in the entire test site were tabulated and area estimates based upon the computer classification were obtained. A separate team of people utilized planimeters and dot grids to determine the area of the various cover types according to the type maps, which had been

Table 1 -- Classification Performance of Major Cover Types in the San Juan Mountain Test Site

Cover Type	No. of Samples[1]	No. of Samples Classified as: Coniferous	Deciduous	Grassland	Barren	Water	Shadow[2]	Percent Correct
Coniferous	9,634	9,110	22	53	21	96	332	94.6
Deciduous	1,475	113	1,286	76	0	0	0	87.2
Grassland	3,677	49	129	2,988	510	0	1	81.2
Barren	35	0	0	1	34	0	0	97.1
Water	1,349	6	0	0	0	1,334	9	98.9
Totals	16,170	9,278	1,437	3,118	565	1,430	342	

Overall Performance = (9,110 + 1,286 + 2,988 + 1,334)/16,170 = 91.2%

[1] Each "sample" is a LANDSAT resolution element. The column labelled "No. of Samples" indicates the total number of resolution elements of the various cover types actually present in the test areas (assuming that each test area contains only a single cover type).

[2] One of the 14 spectral classes that had been defined involved areas of topographic shows, but since this was not an actual cover type, any resolution elements belonging to this spectral class were considered as errors in the classification.

developed from aerial photos using standard photo-interpretation techniques. A random sample of seven quadrangles (totalling about 112,400 hectares) were utilized for the photo-interpretation acreage estimates. Comparison of these two data sets resulted in a correlation coefficient (r) of 0.97. Such a high correlation coefficient indicates that the area estimates obtained by computer analysis of LANDSAT data are in close agreement with the estimates obtained from aerial photos, thereby providing additional confidence in the classification accuracy.

A cost evaluation of this analysis indicated that the total cost (including computer, personnel salaries, etc.) was approximately $0.0025 per hectare (0.1¢ per acre). Since this analysis was done on a medium-speed digital computer programmed for research types of activities, it would appear that in the

future, such analyses could be conducted on special purpose, high speed digital computers in a relatively cost-effective manner.

A second phase in the computer-aided analysis of the satellite data from this area involved mapping forest cover types over a more limited test site, referred to as the Vallecito Intensive Study Site, which covered an area of about 23,000 hectares. The analysis procedures previously described were again utilized in this classification but since a more detailed level of classification was involved, a larger number of spectral classes had to be defined and utilized. In this case, a total of 24 spectral classes were required for the analysis. The computer classification of forest cover types resulted in a map of this intensive study site that qualitatively looked reasonably good, but not as accurate as the map of major cover types. In some cases, individual forest cover types appeared to have been misclassified within the general categories of coniferous or deciduous. This was substantiated by the quantitative test area results, which are shown in Table 2.

Table 2 -- Classification Performance of Forest Cover Types for the Vallecito Intensive Study Site.

Cover Type[1]	No. of Samples	No. of Samples Classified As: Pine	Spruce/Fir	Oak	Aspen	Grassland	Water	Barren	Percent Correct
Pine	1,111	904	169	5	9	3	1	20	81.4
Spruce/Fir	747	254	485	2	6	0	0	0	64.9
Oak	481	8	0	297	95	81	0	0	61.7
Aspen	204	5	0	33	160	6	0	0	78.4
Grassland	242	2	0	6	0	232	0	2	95.9
Water	240	0	0	0	0	0	240	0	100.0
Barren	98	0	0	0	0	6	0	92	93.9
Totals	3,123	1,173	654	343	270	328	241	114	

Overall Performance = (904 + 485 + 297 + 160 + 232 + 240 + 92)/3,123 = 77.2%

[1] Pine = Ponderosa Pine; Spruce/fir = Engelmann spruce, Douglas fir and subalpine fir; Oak = Gambel oak; Aspen = Quaking aspen.

As one might expect, the results shown in Tables 1 and 2 indicate that a better classification performance can be achieved for mapping deciduous and coniferous forest cover and other major cover types than for mapping individual forest cover types. However, I believe that these results for computer-aided analysis of forest cover types are reasonably good, particularly when the vegetative and topographic complexity of the test site is considered. Additional statistical analysis of the LANDSAT data showed that the spectral response was significantly influenced by differences in stand density, elevation, aspect, and slope, as well as the differences between the forest species.

A third study on this test site utilized both SKYLAB and LANDSAT multispectral scanner data obtained on the same day (June 5, 1973). The S-192 MSS

system on SKYLAB obtained data in thirteen wavelength bands, which included wavelengths in the middle and thermal infrared portions of the spectrum. In addition to this greater spectral range, the spectral resolution of the SKYLAB wavelength bands was better than the LANDSAT bands. Unfortunately, the quality (signal/noise ratio) of this SKYLAB data was not as good as the LANDSAT data, and consequently the real value of the improved spectral range and spectral resolution of the SKYLAB scanner was not obvious in the results obtained (Hoffer, 1975).

Although there was a difference in the orbital paths of LANDSAT and SKYLAB of about 58°, the geometric correction, scaling, and overlay procedures that we had developed were able to successfully correct for this difference, and the data were digitally registered with a high degree of accuracy. This procedure allowed a group of test areas to be defined that were common to both sets of satellite data, thereby enabling an precise comparison of the results. These test areas (shown previously in Figure 3) included 2400 resolution elements.

The SKYLAB and LANDSAT scanner data were classified into both major cover types and individual forest cover types, again utilizing the techniques described previously. The quantitative test area results showed that major cover types, including coniferous and deciduous forest categories, had been classified with an overall accuracy of 85.7% using the LANDSAT data, and 85.0% with the SKYLAB data.[1/] The individual forest cover types had a classification accuracy of 68.4% with the LANDSAT data, and 71.0% for the SKYLAB data. The quantitative evaluation of acreage estimates (shown in Figure 4) resulted in a correlation coefficient of 0.929. These results thus provide additional evidence that forest cover can be mapped with a reasonable degree of accuracy using these computer-aided analysis techniques.

In a final phase of this analysis, topographic data (elevation, slope, and aspect) were also digitially registered or overlayed onto the combined SKYLAB and LANDSAT data. Incorporation of this elevation data into the analysis sequence caused an improvement in classification performance of over 10% for both spruce-fir and aspen forest cover types. This result indicates that when two cover types are spectrally similar but occur in different elevation zones, the combination of elevation plus multispectral scanner data will allow significant improvements to be obtained in accurately differentiating and mapping such cover types

I believe that the results described above are significant because they involve several rather than a single analysis sequence, a relatively large test site containing a complex mixture of vegetative cover and topography was involved, the results are expressed quantitatively and include a relatively large test data set, and because there was a reasonable degree of consistency in the results obtained in the different analysis sequences. In each case, major cover types (including deciduous and coniferous forest categories) were identified and mapped with approximately 85% accuracy, the individual forest cover types had about 70-75% accuracy, and acreage comparisons resulted in correlation coefficients of 0.93-0.97.

[1/] All four wavelength bands of LANDSAT data (0.5-0.6, 0.6-0.7, 0.7-0.8, and 0.8-1.1μm) were used. The "Best 4" bands of SKYLAB data used for this analysis were defined as the 0.46-0.51, 0.78-0.88, 1.09-1.19, and 1.55-1.75μm wavelength bands.

SUMMARY AND CONCLUSIONS

Geometric correction and scaling of multispectral satellite data using digital techniques allows accurate 1:24,000 line-printer outputs to be obtained. A "modified clustering technique" has proven to be superior to both the "supervised" and "non-supervised" (or clustering) techniques that have been used in the past to define training statistics for computer-aided analysis of multispectral scanner data. A combination of three different techniques, including qualitative evaluation, quantitative test field performance, and acreage comparisons, provide the best overall approach for evaluation of the results obtained by computer-aided analysis of satellite data.

Classification of both LANDSAT and SKYLAB satellite data showed that coniferous and deciduous forest cover (as well as other major cover types) could be identified and mapped with reasonable accuracy (85%) using these techniques. In spite of the vegetative and topographic complexity of the test site, individual forest cover types could be classified and mapped with about 70% accuracy. Acreage estimates of forest cover obtained by computer-aided analysis of satellite data were highly correlated ($r = 0.93$-0.97) with acreage estimates obtained by standard photo-interpretation techniques using aerial photography.

The application of computer-aided analysis techniques to multispectral scanner data from satellite altitudes has been proven feasible. These results indicate a significant potential for mapping and tabulating many of our natural resources over large geographic areas in a quantitative, rapid, and cost-effective manner.

LITERATURE CITED

Anderson, J. R., E. E. Hardy and J. T. Roach. 1972. "A Land-Use Classification System for Use With Remote-Sensor Data". Geological Survey Circular 671, U.S. Geological Survey, Washington, D.C. 16pp.

Fleming, M., J. Berkebile, and R. Hoffer. 1975. "Computer-Aided Analysis of LANDSAT-1 MSS Data: A Comparison of Three Approaches Including the 'Modified Clustering' Approach". Proceedings of the Symposium on Machine Processing of Remotely Sensed Data, Purdue University, West Lafayette, Indiana, June 3-5, 1975. pp. 1B-54 to 1B-61.

Hoffer, R. M. and Staff. 1975. "Computer-Aided Analysis of SKYLAB Multispectral Scanner Data in Mountainous Terrain for Land Use, Forestry, Water Resource, and Geologic Applications". SKYLAB Final Report. LARS Information Note 121275, Laboratory for Applications of Remote Sensing, Purdue University, West Lafayette, Indiana. 381pp.